Detecção de Vazamentos no Fluxo de Água com Aplicação de Machine Learning

Alexandre Bedin[1], William Moraes da Silva[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul
Farroupilha – RS – Brasil

alexandre.bedin@aluno.farroupilha.ifrs.edu.br,
william.silva@farroupilha.ifrs.edu.br

Resumo. *A perda de água em sistemas de distribuição representa um dos principais desafios para a sustentabilidade ambiental e a eficiência operacional dos serviços de saneamento. Diante desse problema, este estudo propõe o desenvolvimento de um sistema de detecção de vazamentos baseado em técnicas de inteligência artificial, com foco no aprendizado de máquina. O principal objetivo da pesquisa foi aplicar o algoritmo Random Forest para identificar padrões anômalos no fluxo de água, possibilitando a detecção precoce de vazamentos. Para isso, foi construída uma base de dados sintética que simula medições de nível, pressão e vazão de um reservatório, permitindo o treinamento e a validação do modelo. Os resultados obtidos foram expressivos, alcançando acurácia global de 99,65%, com alta precisão (96%) e recall (93%), o que demonstra a capacidade do modelo em identificar de forma confiável as anomalias associadas a possíveis vazamentos. Conclui-se que a integração de aprendizado de máquina em sistemas de monitoramento hídrico é uma alternativa viável para aprimorar a gestão dos recursos e reduzir perdas, além de abrir caminho para futuras aplicações em ambientes reais e integrados com sensores.*

Abstract. *Water loss in distribution systems represents one of the main challenges for environmental sustainability and the operational efficiency of sanitation services. To address this issue, this study proposes the development of a leak detection system based on artificial intelligence techniques, with a focus on machine learning. The main objective of the research was to apply the Random Forest algorithm to identify anomalous patterns in water flow, enabling early leak detection. To this end, a synthetic database was constructed that simulates reservoir level, pressure, and flow measurements, enabling model training and validation. The results were highly promising, achieving an overall accuracy of 99.65%, with high precision (96%) and recall (93%), demonstrating the model's ability to reliably identify anomalies associated with potential leaks. The conclusion is that integrating machine learning into water monitoring systems is a viable alternative for improving resource management and reducing losses, in addition to paving the way for future applications in real-world environments integrated with sensors.*

1.Introdução

A crescente demanda por eficiência no uso de recursos naturais, aliada à necessidade de reduzir perdas e desperdícios, tem impulsionado a aplicação de tecnologias avançadas no monitoramento e gerenciamento de sistemas. No segmento de gestão hídrica, o monitoramento em tempo real do fluxo de água e a detecção de vazamentos são desafios críticos enfrentados por diversos setores, como indústrias, saneamento básico e agricultura. No Brasil, o volume de água produzido em 2022 atingiu 17,9 bilhões de m³, enquanto o volume consumido foi de 11,7 bilhões de m³. (SANSA, 2022)

O desenvolvimento de um modelo de dados representa um avanço significativo em relação aos métodos tradicionais, que dependem de inspeções físicas e medições pontuais através de sensores. Embora a medição contínua seja fundamental, a amostra do SNSA sobre 2022 aponta para a persistência de gargalos na medição e na gestão da informação. Estudos semelhantes indicam que modelos baseados em *machine learning* têm potencial para alcançar altas taxas de precisão, mas enfrentam desafios em sua implementação prática e na adaptação a diferentes cenários operacionais. Este trabalho busca preencher essas lacunas, com o desenvolvimento de um modelo de aprendizado de máquina. (Nusda, 2023; Flores et al., 2021)

A área de aprendizado de máquina (*machine learning*) se dedica à criação de sistemas e algoritmos que permitem que computadores adquiram conhecimento a partir de dados, sem a necessidade de uma programação explícita para cada tarefa específica. criando cenários antecipados em amplas áreas. Com essa capacidade, eles passam a realizar tarefas de forma independente, evoluindo continuamente conforme recebem novas informações e experiências. (Silva, 2025)

Este trabalho aborda a problemática relacionada à identificação de vazamentos em sistemas de distribuição de água, que impactam diretamente a sustentabilidade ambiental e a eficiência econômica das operações. Diante dessa situação, buscou-se desenvolver um modelo de aprendizado de máquina para detecção de padrões anômalos no fluxo de água, indicando possíveis vazamentos. Esta pesquisa é relevante para a gestão sustentável de recursos hídricos, pois até 37,8% da água pode ser perdida em sistemas de distribuição. Esse índice nacional esconde uma realidade ainda mais grave em algumas regiões, como o Norte e Nordeste, que registraram índices de perdas de 46,9% e 46,7%, respectivamente. Além disso, a situação é alarmante em algumas capitais, onde as perdas podem chegar a 77,3%. Tais perdas estão associadas a fatores como a baixa medição dos grandes volumes de água macromedição, que em 2022 era de 72,3% no país e a falta de investimentos em redes e operações. (SANSA, 2022)

Este artigo está organizado da seguinte forma: na seção 2 são abordados os objetivos geral e específicos; na seção 3 estão detalhados os procedimentos metodológicos adotados, apresentando a construção de uma base de dados sintética e as tecnologias selecionadas para o estudo; na seção 4 está o desenvolvimento de um algoritmo de *Random Forest* para identificar possíveis casos de vazamentos; por fim, os resultados obtidos são analisados na seção 5.

2.Objetivos

2.1. Objetivo geral

Desenvolver e avaliar um modelo de Random Forest para detecção de vazamentos em sistemas de distribuição de reservatório de água.

2.2. Objetivos específicos

- Construir uma base de dados sintética representando o fluxo de água em sistemas de distribuição, e implementar um algoritmo de Random Forest para identificar padrões anômalos.
- Realizar ajustes e treinamento no modelo para identificar os melhores resultados.
- Avaliar os resultados obtidos através de análises de métricas específicas.

3. Referencial Teórico

3.1. Água, Disponibilidades e Usos

A escassez da água representa um dos principais desafios para o planejamento de longo prazo de diversas regiões do Brasil. A água é um recurso fundamental para a agricultura, as atividades industriais, a geração de energia e a perpetuidade da espécie humana, uma vez que sem água, não há vida. Apesar de tamanha importância, uma série de atores governamentais negligenciam as suas responsabilidades para a preservação ambiental. Nesse contexto, tem-se que a qualidade ambiental e a preservação das águas são responsabilidade comum de todos os entes federativos, sendo que é nas cidades onde os problemas ambientais atingem maior amplitude. (Braz e; Longo, 2021)

A água doce representa apenas 2,5% da disponibilidade hídrica mundial, sendo a maior parte de difícil exploração. Nesse cenário, os sistemas de abastecimento de água assumem papel essencial, englobando captação, transporte, tratamento, armazenamento e distribuição para diferentes setores da sociedade. No Brasil, o Sistema Nacional de Informações sobre Saneamento (SNSA) fornece dados que permitem compreender o panorama nacional. Apesar de o país concentrar cerca de 12% da água doce mundial, sua distribuição é desigual: a região Norte concentra a maior parte dos recursos, mas possui baixa densidade populacional, enquanto Sudeste e Nordeste, com 69% da população, dispõem de menos de 10% da água disponível para consumo. (SANSA, 2022)

Se a água doce é indispensável para diversas atividades, em especial para a agricultura irrigada e para os processos industriais, também não há de se esquecer que é um recurso crucial para a dessedentação humana e animal, bem como para a conservação da natureza. Assim, a água deve ser tratada como um elemento diretamente ligado à sustentabilidade da sociedade. (Jacobi, Buckeridge e Ribeiro, 2021)

3.2. Sistemas de Abastecimento e Redes de Distribuição de Água

Para que a água possa chegar a todos os lugares, faz-se imprescindível uma adequada rede de distribuição de água, composta por tubulações, bombas, acessórios, reservatórios e demais equipamentos que permitam atender, dentro de condições sanitárias, de vazão e pressão, aos pontos de consumo de uma cidade. Para que essa rede possa cumprir satisfatoriamente seu objetivo, é necessário um correto dimensionamento, que deve considerar fatores como a topografia, o tipo de manancial disponível, a população a ser atendida e sua estimativa de crescimento, de modo que o sistema seja capaz de atender adequadamente a população por muitos anos. Os mananciais de abastecimento podem incluir poços profundos e/ou rasos, lagos, rios e reservatórios formados por barragens. (Tsutiya, 2006; Porto, 2006)

Existem perdas em redes de distribuição de água que podem ser aparentes ou comerciais. Essas perdas podem ocorrer devido a ligações clandestinas, falhas em hidrômetros, contabilizações erradas do consumo de água e outros usos não autorizados. Além disso, há perdas reais, causadas por vazamentos oriundos de excessos de pressão, trincas nas tubulações e outras falhas. (Kanakoudis e Muhammetoglu, 2014)

Kanakoudis e Muhammetoglu (2014) mencionam perdas na ordem de 45 bilhões de m³ em todo o mundo, sendo que a redução pela metade dessas perdas seria capaz de abastecer cerca de 200 milhões de pessoas sem a necessidade de exploração de novos mananciais, o que representa quase o total da população brasileira.

As perdas representam um significativo prejuízo financeiro para as companhias, na ordem de US$ 14 bilhões ao ano em todo o mundo, sendo particularmente crítico em países em desenvolvimento, pois limitam a capacidade das companhias de investir em melhorias e expansão do sistema. O projeto de rede de distribuição de água deveria balancear o custo de implantação, que geralmente é baixo, com o custo de manutenção, que pode ser elevado devido à alta frequência de intervenções. (Pinnto et al., 2017)

As perdas reais afetam diversos aspectos da operação das redes de distribuição de água, como a qualidade do serviço, o consumo energético e de insumos para tratamento, perdas econômicas e até mesmo riscos sanitários. Elas estão diretamente relacionadas ao envelhecimento das tubulações e ao excesso de pressão na rede, que causam rupturas. Dada a importância do problema, as companhias de saneamento empregam esforços para reduzir suas perdas reais por meio do controle de pressões. (Giustolisi, Ridolfi e Simone, 2019)

A Figura 1 é um exemplo de como é feito a distribuição de água e suas composições como as tubulações, bombas, reservatórios e estação de tratamento (Porto, 2006).

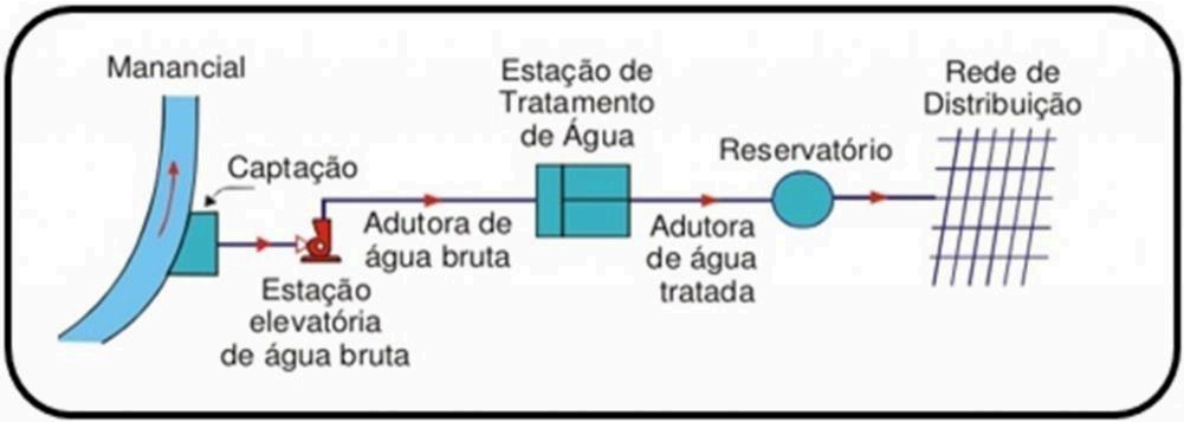

Figura 1. Rede de distribuição de água
Fonte: Porto (2006)

Para que a rede de distribuição de água funcione de maneira eficiente, é essencial que seu dimensionamento seja realizado corretamente, levando em conta fatores como a topografia, o tipo de manancial disponível, a população atendida e suas projeções de crescimento. Assim, o sistema será capaz de atender às necessidades da população por um longo período. Há perdas reais relacionadas a vazamentos causados por excesso de pressão, trincas em tubulações, entre outros fatores. Kanakoudis e Muhammetoglu, 2014; Tsutiya, 2006)

A medição contínua desempenha um papel crucial no controle de sistemas de abastecimento de água, especialmente no gerenciamento de perdas. Ela abrange desde a captação até a distribuição e o consumo, permitindo identificar diferenças entre o volume de água produzido e o consumido. A macromedição, realizada em pontos estratégicos da rede, como saídas das Estações de Tratamento de Água (ETAs), mede grandes volumes de água, enquanto a micromedição, ou hidrometração, ocorre no atendimento ao consumidor final por meio de hidrômetros (SANSA, 2022)

O conjunto de medições de grandes volumes de água realizado no sistema público de abastecimento é apresentado na Figura 2, abrangendo desde a captação até as extremidades da rede. Essas medições auxiliam na identificação dos volumes efetivamente distribuídos e das perdas de água no sistema. Observa-se que a região Sudeste apresenta o maior volume produzido, reflexo de sua alta densidade populacional e demanda concentrada, enquanto a região Norte registra o menor volume, consequência direta da menor ocupação urbana e da ampla disponibilidade de recursos hídricos. Essa disparidade evidencia a necessidade de estratégias diferenciadas de gestão e monitoramento, considerando as particularidades de cada macrorregião do país.

Figura 2. Volume produzido em m³ por macrorregião
Fonte: SANSA (2022)

Note-se na Figura 3 que o acompanhamento do consumo médio de água é uma ferramenta essencial para o controle operacional e o planejamento dos serviços de abastecimento. Ele auxilia no dimensionamento de sistemas em municípios com crescimento populacional e no controle do aumento do consumo em regiões com recursos hídricos limitados. O cálculo do índice de consumo médio diário per capita baseia-se no volume de água consumido, excluindo a quantidade transferida para outros distribuidores. O resultado é dividido pela média aritmética da população atendida nos

últimos dois anos. O controle do consumo médio de água é uma ferramenta fundamental para planejar, administrar e operar serviços de abastecimento. Ele auxilia na concepção de sistemas em cidades em expansão e na regulação do consumo em áreas com recursos hídricos escassos. (SANSA, 2022)

Figura 3. Volume de litros por habitantes nas macrorregiões.
Fonte: SANSA (2022)

Em 2022, o consumo médio de água per capita no Brasil foi de 148,2 litros por habitante ao dia (l/hab.dia), representando uma redução de 3,3% em relação aos 153,3 l/hab.dia registrados em 2021. Entre as macrorregiões, Sul, Norte, Centro-Oeste e

Sudeste apresentaram índices acima da média nacional, com 149,8, 151,2, 153,5 e 159,9 l/hab.dia, respectivamente. O menor consumo foi registrado no Nordeste, com 121,4 l/hab.dia. Quanto à abrangência dos serviços, o consumo diário per capita variou entre 110,8 l/hab.dia, na prestação Microrregional, e 193,6 l/hab.dia. (SANSA, 2022)

3.3. Perdas de Água em Sistemas de Distribuição

As perdas são inevitáveis em qualquer sistema de abastecimento de água. Esse assunto tem ganhado mais destaque nas últimas décadas devido ao aumento da frequência de eventos de escassez hídrica e do risco crescente de contaminação da água tratada, uma vez que as perdas podem reduzir a pressão na rede de distribuição. Além disso, do ponto de vista econômico e financeiro, há questões relacionadas aos custos da energia elétrica utilizada na produção e distribuição da água tratada, bem como ao desperdício de recursos naturais e operacionais. (SANSA, 2022)

Os custos resultantes das perdas são repassados ao consumidor final. Perdas elevadas de água podem comprometer o direito humano de acesso à água potável, reconhecido pela Organização das Nações Unidas, e o princípio da universalização do acesso aos serviços de saneamento básico, estabelecido pela Lei nº 11.445/2007. Nesse contexto, são essenciais ações estruturais e programas contínuos e eficazes de avaliação, controle e redução de perdas. Paralelamente, às ações do SNSA, que publicam a situação das perdas de água dos prestadores de serviço no Brasil, ajudam as agências reguladoras em suas atividades de fiscalização e monitoramento. (SANSA, 2022)

Dados do SANSA (2022) indicam um índice de perdas na distribuição de água de 37,8%. Em relação a 2021, houve uma redução de 2,9 pontos percentuais no índice de perdas, que vinha apresentando um aumento contínuo desde 2015. Esse percentual representa a fração do volume de água disponibilizado que não foi faturado por não ter sido contabilizado como volume utilizado pelos consumidores, seja por vazamentos, falhas nos sistemas de medição ou ligações clandestinas. Em termos quantitativos, o índice significa que, de cada 100 litros disponibilizados pelos prestadores de serviços, apenas 62,2 litros são contabilizados como utilizados pelos consumidores. O índice de perdas por macrorregiões é demonstrado na Figura 4, onde a grande maioria dos estados teve uma leve diminuição nas perdas, mas o desperdício ainda se mantém significativamente alta.

PERDA DE DISTRIBUIÇÃO DE ÁGUA MÉDIO POR ESTADO EM 2022 E 2021				
Macrorregião	Estados	Indice de perdas de água (%)		
		2022 (%)	2021 (%)	Variação absoluta
Norte	Acre	66,6	74,4	-7,8
	Amapá	71,1	74,8	-3,7
	Amazonas	50,9	53	-2,1
	Pará	34,6	37,4	-2,8
	Rondônia	59,8	61,4	-1,6
	Roraima	59,4	64	-4,6
	Tocantins	39,7	42,5	-2,8
Nordeste	Alagoas	39,7	46,9	-7,2
	Bahia	42,5	39,7	2,8
	Ceará	44,4	45,2	-0,8
	Maranhão	57	59,2	-2,2
	Paraíba	37	35,4	1,6
	Pernambuco	48,5	46	2,5
	Piauí	47,5	45,3	2,2
	Rio Grande do Norte	49,3	52,2	-2,9
	Sergipe	57,6	48,4	9,2
Sudeste	Espírito Santo	37,6	38,8	-1,2
	Minas Gerais	36,8	37,5	-0,7
	Rio de Janeiro	32	45	-13
	São Paulo	34,1	34,5	-0,4
Sul	Paraná	35,1	33,8	1,3
	Rio Grande do Sul	39,5	41,6	-2,1
	Santa Catarina	34,7	34,1	0,6
Centro-Oeste	Distrito Federal	33,8	35,1	-1,3
	Goiás	28,3	28,5	-0,2
	Mato Grosso	45,4	48,4	-3
	Mato Grosso do Sul	33,2	33,4	-0,2

Figura 4. Perdas em porcentagem nos Estados Brasileiros entre 2021 e 2022. Fonte: SANSA (2022)

3.4. Como Diminuir as Perdas

O volume de água perdida é um indicador fundamental da eficiência das empresas de abastecimento, pois reflete a qualidade do planejamento, construção e manutenção do sistema. Altos volumes anuais de perda de água, especialmente se em crescimento, apontam para uma ineficiência no planejamento e construção, além de demonstrarem deficiências na manutenção e operação do sistema. A redução das perdas de 45% para 25% nas empresas brasileiras de saneamento poderia liberar recursos financeiros na ordem de 1 bilhão de reais por ano. A deficiência dos sistemas de abastecimento de água é consequência de uma falta de planejamento e manutenção adequada, combinada com a escassez de recursos financeiros. Com o tempo, esses sistemas se deterioram, seja de forma natural ou acelerada, resultando em problemas operacionais que diminuem a

qualidade dos serviços prestados e aumentam os custos operacionais, que são repassados aos consumidores através de tarifas. (Favretto et al., 2016)

É essencial adotar uma metodologia de apoio à decisão multicritério para auxiliar gestores na avaliação de ações voltadas ao abastecimento de água, considerando critérios financeiros, técnicos, ambientais e sociais, bem como as particularidades de cada sistema. Essa abordagem permite escolher medidas preventivas ou corretivas que garantam uma distribuição equilibrada, sustentável e adequada à realidade das concessionárias e às necessidades dos consumidores. (Al-Rashdan et al., 1999)

O uso de m*achine learning* para controlar ou prevenir vazamentos de água é justificado pela capacidade desses modelos de detectar rapidamente anomalias com alta precisão a partir de dados de pressão, vazão e outros parâmetros coletados em tempo real por sensores e sistemas. Modelos como Redes Neurais Artificiais podem aprender a diferenciar condições normais de anômalas, mesmo com conjuntos de dados desbalanceados, e operam com sensores de baixo custo e baixa taxa de amostragem. Além disso, a integração com tecnologias de Internet das Coisas (IoT) e medidores inteligentes permite a detecção em tempo real, reduzindo perdas, custos operacionais e impactos socioeconômicos. (Jesus, 2023)

3.5. Inteligência Artificial

Em 1956, John McCarthy organizou no Dartmouth College um seminário que marcou o nascimento da inteligência artificial (IA). Com apoio de Marvin Minsky, Claude Shannon e Nathaniel Rochester, o encontro reuniu pesquisadores interessados em máquinas capazes de simular aspectos da inteligência humana. O objetivo era explorar como computadores poderiam aprender, usar linguagem e resolver problemas como os humanos. Esse evento estabeleceu as bases conceituais da IA moderna. (Russell e Novig, 2011)

A inteligência artificial já está presente em diversas áreas, como veículos autônomos, capazes de dirigir sozinhos. Também é usada em sistemas de reconhecimento de voz, no planejamento autônomo de missões espaciais, e em jogos, como o DEEP BLUE, que derrotou Kasparov no xadrez. Além disso, algoritmos de IA combatem spams e otimizam planejamentos logísticos militares. Essas aplicações mostram que a IA é resultado de ciência e engenharia, não de ficção científica. (Russell e Novig, 2011)

3.6. *Machine Learning*

O aprendizado de máquina pode ser entendido como um processo automatizado capaz de identificar padrões em conjuntos de dados. Para desenvolver os modelos aplicados em análises preditivas, utiliza-se o aprendizado de máquina supervisionado. Essas técnicas supervisionadas permitem que o sistema aprenda, de forma automática, um modelo que descreve a relação entre diversas variáveis descritivas e uma variável-alvo, com base em um conjunto de dados históricos. (Kelleher, MacNamee e D'Arcy, 2015)

A Figura 5 ilustra um exemplo onde Kelleher, MacNamee e D'Arcy (2015)

utilizam em duas etapas o aprendizado de máquina. Na primeira etapa, um algoritmo de aprendizado recebe um conjunto de dados históricos contendo características descritivas e uma variável alvo, aprendendo as relações entre elas para gerar um modelo de predição. Na segunda etapa, esse modelo treinado é utilizado para realizar previsões sobre novos dados, aplicando o conhecimento adquirido a situações inéditas. Assim, o processo envolve primeiro o aprendizado a partir de exemplos conhecidos e, em seguida, a aplicação desse aprendizado para prever resultados desconhecidos.

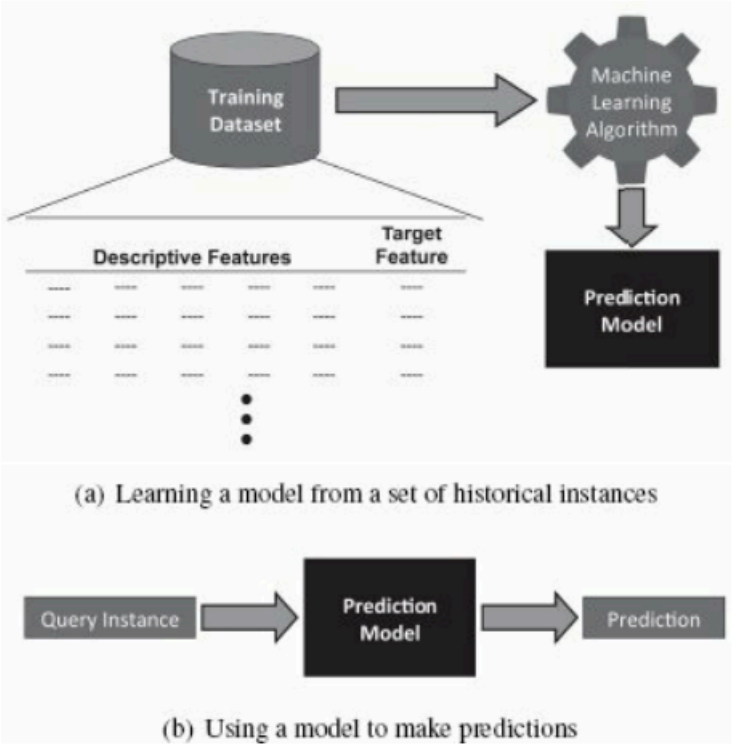

Figura 5. Etapas no aprendizado de máquina.
Fonte: Kelleher, MacNamee e D'Arcy (2015)

Os algoritmos de aprendizado de máquina procuram identificar o modelo que melhor descreve a relação entre as variáveis de um conjunto de dados. No entanto, buscar apenas consistência com os dados pode levar a erros, pois eles podem conter ruídos e representar apenas uma amostra parcial da realidade. Assim, o aprendizado de máquina é considerado um problema mal definido, já que não é possível determinar uma única solução com base apenas nas informações disponíveis. (Kelleher; MacNamee; D'Arcy, 2015)

3.7. *Random Forest*

O algoritmo de *Random Forest* (Floresta Randômica) é uma técnica de aprendizado de máquina que combina várias árvores de decisão para reduzir a correlação entre os dados e melhorar a precisão das previsões. Ele utiliza amostras e variáveis escolhidas aleatoriamente, o que diminui o risco de overfitting e permite processamento paralelo, tornando o método mais rápido e eficiente ao lidar com grandes volumes de dados. (Hasan, Ali e Amani, 2025).

Árvores de decisão são modelos hierárquicos usados em aprendizado supervisionado para classificar dados e prever resultados com base em atributos. Elas começam com o nó raiz, que representa o conjunto completo de dados e é dividido em subnós por meio do processo de divisão, baseado em critérios como ganho de informação. Os nós de decisão representam os pontos de escolha, enquanto os nós folha indicam os resultados finais, onde não há mais divisões possíveis. Para evitar complexidade excessiva e melhorar a generalização, aplica-se a poda, que remove ramificações irrelevantes ou redundantes. Árvores de decisão podem ser binárias, com cada nó gerando dois subnós, ou baseadas em atributos, onde os dados são segmentados conforme características específicas. Essa estrutura reflete o raciocínio humano, tornando o modelo intuitivo e fácil de interpretar. (Hasan, Ali e Amani, 2025)

Este é um algoritmo versátil, aplicável a problemas de classificação e regressão, que oferece bons resultados mesmo sem ajustes complexos e ajuda a reduzir o overfitting ao combinar várias árvores de decisão. Porém, sua principal desvantagem é o tempo elevado de treinamento e a lentidão nas previsões, o que pode torná-lo menos eficiente em aplicações que exigem respostas em tempo real. (Hasan, Ali e Amani 2025)

3.8. Métricas de avaliação

As métricas são usadas para avaliar o desempenho de um classificador, mostrando como medir sua capacidade de prever corretamente os rótulos das amostras em cenários equilibrados ou desbalanceados. No Quadro 1 são apresentadas medidas como acurácia, precisão, *recall*, especificidade, *F1-Score*, destacando que a acurácia, embora específica, também pode indicar o desempenho geral do modelo. (Han, Pei e Tong, 2023)

Quadro 1. Métricas de avaliação.

Medida	Fórmula	Descrição
Acurácia (accuracy, taxa de reconhecimento)	$\frac{TP+TN}{P+N}$	Mede a proporção de previsões corretas (positivas e negativas) em relação ao total de amostras.
Sensibilidade (sensitivity, taxa de verdadeiros positivos, recall)	$\frac{TP}{TP+FN}$	Mede a capacidade do modelo de identificar corretamente as instâncias positivas.
Precisão (precision)	$\frac{TP}{TP+FP}$	Mede a proporção de previsões positivas que realmente são positivas.

F, F1 ou F-score (média harmônica entre precisão e sensibilidade)	$\frac{2 \times \text{precisão} \times \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}}$	Combina precisão e sensibilidade em uma única métrica, equilibrando ambas.

Fonte: (Han, Pei e Tong, 2023)

A acurácia indica a proporção de amostras corretamente classificadas por um modelo em um conjunto de teste, refletindo sua capacidade de generalização. Quando essa avaliação é feita utilizando o próprio conjunto de treinamento, obtém-se o chamado erro de re-substituição, que costuma ser irrealisticamente baixo, já que o modelo é testado em dados que já conhece. Por isso, esse tipo de medida tende a superestimar o desempenho real do classificador, não representando fielmente sua eficácia em novos dados. (Han, Pei e Tong, 2023)

O Recall é uma métrica fundamental para avaliar o desempenho de classificadores em problemas de classificação binária. Ele mede a proporção de casos positivos corretamente identificados pelo modelo em relação ao total de casos positivos existentes, conforme representado na matriz de confusão. Em outras palavras, o Recall indica quão eficaz o modelo é em detectar todos os exemplos da classe de interesse, sendo especialmente importante em contextos nos quais falsos negativos são críticos, o Recall complementa a métrica de Precisão, permitindo uma avaliação mais completa da capacidade do classificador de identificar corretamente os casos positivos. (Han, Pei e Tong, 2023)

O F1-Score integra as métricas de Recall e Precisão em uma única medida harmônica, oferecendo uma avaliação conjunta da habilidade do modelo em identificar corretamente os casos relevantes e da confiança nas suas previsões. Essa métrica se mostra especialmente valiosa em contextos com classes desbalanceadas, onde a Acurácia isolada pode levar a interpretações equivocadas sobre o real desempenho do modelo.(Han, Pei e Tong, 2023).

3.9. Estudos prévios

Diversas pesquisas recentes têm investigado a aplicação de técnicas de aprendizado de máquina na detecção de vazamentos e na otimização de sistemas de distribuição de água. Esses estudos visam aprimorar a identificação de padrões anômalos, minimizar perdas e fornecer suporte à tomada de decisões em tempo real. A análise da literatura existente permite identificar as metodologias mais recorrentes e reconhecer os principais desafios ainda enfrentados, formando uma base sólida para a criação do modelo proposto neste trabalho.

O estudo de Gouveia (2022) analisou o uso de técnicas de aprendizado de máquina, incluindo modelos tradicionais e *ensemble learning*, para aprimorar a gestão de ativos e o controle de perdas em sistemas de abastecimento de água. Ele buscou comparar modelos preditivos, identificar fatores que influenciam falhas na rede, avaliar

o impacto de variáveis hidráulicas e aplicar métodos avançados para melhorar a precisão e apoiar a tomada de decisão na redução de perdas de água.

Aplicando diversos modelos de aprendizado de máquina por classificação, incluindo *Linear SVM* (*Support Vector Machine Linear*), *Radial SVM* (*Support Vector Machine Linear*), Regressão Logística, KNN (*K-Nearest Neighbors*), *Decision Tree*, *Naive Bayes* e *Random Forest*, além de ensemble learning models como *Bagged KNN*, *Bagged Decision Tree*, *Adaboost*, *Gradient Boosting* e *XGBoost* para prever vazamentos em ramais de água. Todos os modelos foram treinados usando validação cruzada com 10 pastas. A abordagem múltipla permitiu comparar a performance dos métodos, buscando maximizar a acurácia e melhorar a qualidade das decisões no controle de perdas de água.

Em contrapartida, o estudo de Jesus (2023) teve como objetivo avaliar a viabilidade de um modelo híbrido que combina Máquinas de Vetores de Suporte para regressão e Redes Neurais Artificiais na previsão da demanda de água nos reservatórios da região metropolitana de Salvador. Utilizando dados históricos de consumo e informações meteorológicas, o modelo foi desenvolvido com base em uma metodologia previamente estudada e se mostrou viável, embora não igualmente eficaz para todos os reservatórios. Além disso, modelos individuais como a Rede Neural Perceptron Multicamadas, a Máquina de Vetores de Suporte e o Modelo Autorregressivo de Média Móvel também apresentaram bom desempenho, com erro percentual médio absoluto de cerca de 3%.

Este estudo propõe uma abordagem para detecção de vazamentos e otimização de sistemas de abastecimento de água, utilizando aprendizado de máquina. Ao contrário de trabalhos anteriores que empregaram bases de dados reais e modelos híbridos complexos, a pesquisa desenvolve uma base sintética e aplica o algoritmo Random Forest para classificação supervisionada, permitindo simulações em cenários controlados. A revisão da literatura fundamenta o modelo proposto, destacando avanços, metodologias recorrentes, com o objetivo de oferecer uma solução prática, e adaptável a sistemas reais de monitoramento hídrico.

4. Procedimentos Metodológicos

A metodologia é o alicerce de qualquer trabalho científico, proporcionando clareza sobre como os objetivos propostos foram alcançados. Este processo envolve etapas estruturadas de gerar dados, pré-processamento, modelagem e validação. (Creswell e Creswell, 2021)

A implementação de ferramentas para aprendizado de máquina supervisionado e análise estatística em Python foram obtidas com o auxílio da função *classification_report* da biblioteca *Scikit-learn.* (Pedregosa et al., 2011)

Desta forma, optou-se por uma abordagem quantitativa e experimental, que se justifica pela análise de dados numéricos em larga escala, aplicando métodos de aprendizado de máquina para identificar anomalias.

Um ciclo iterativo, composto pela definição do problema, desenvolvimento incremental e validação contínua foi seguido para garantir a qualidade do sistema desenvolvido. Os objetivos específicos incluem implementar um pipeline de *machine learning* com a técnica *Isolation Forest* para identificar anomalias, criar um banco de dados relacional para armazenar dados de sensores, utilizando Linguagem de Consulta Estruturada, desenvolver uma interface gráfica interativa com suporte a visualizações de dados, e avaliar o desempenho do modelo em cenários simulados. A estrutura metodológica foi planejada em quatro etapas principais, inspirada nas recomendações de (Russell e Novig, 2011).

Para o desenvolvimento deste estudo, foi gerada uma base de dados sintética inspirada em padrões reais de comportamento de sistemas de abastecimento de água, simulando medições de nível, pressão e vazão. A partir dessa base, implementou-se o algoritmo de aprendizado de máquina Random Forest para identificação de anomalias associadas a possíveis vazamentos. Todo o processo foi realizado na plataforma Google Colab, utilizando a linguagem de programação Python e as bibliotecas Pandas, NumPy, Matplotlib e Scikit-learn, que forneceram suporte para a manipulação dos dados, modelagem, treinamento e avaliação do desempenho do modelo.

4.1. Geração de Dados

O código cria uma simulação realista de medições operacionais de nível água, pressão e vazão, permitindo realizar testes, análises e treinamentos de modelos de aprendizado de máquina sem depender de dados reais e privados. Para criar uma base de dados referentes às leituras de um reservatório (nível, pressão e vazão), foi utilizada a linguagem de programação Python e as bibliotecas pandas, numpy e datetime.

Primeiramente, foram definidos os parâmetros que determinam o intervalo e o período de geração dos dados. As leituras foram configuradas para ocorrer a cada 5 minutos, representando medições periódicas e contínuas, enquanto o período de simulação abrange dois anos, considerando a data atual (*datetime.now()*) e retrocedendo 2 anos. O nível máximo do reservatório foi fixado em 10 metros, servindo como referência para o cálculo do percentual de nível. Esses parâmetros possibilitam a criação de um conjunto de dados coerente e representativo de medições reais obtidas por sensores de um reservatório.

Em seguida, foi criada uma série temporal utilizando a função *date_range()* da biblioteca pandas, responsável por gerar uma sequência contínua de datas e horários espaçados a cada 5 minutos entre a data inicial e final definidas. O resultado dessa operação foi armazenado na variável *datas*, que serviu como eixo temporal do *data frame*, permitindo organizar e associar corretamente cada medição simulada ao seu respectivo instante de tempo. O número total de registros (n) foi determinado a partir do tamanho dessa sequência, representando o total de leituras geradas.

O nível, em metros, foi gerado com distribuição normal de média 8,0 m e desvio padrão de 0,3m; a pressão, em bar, com média 3,5 e desvio padrão de 0,15; e a vazão, em metros cúbicos por hora, com média 25,0 e desvio padrão de 5,0. Para isso, utilizou-se o método *random.normal()* da biblioteca numpy, criando as distribuições

gaussianas (distribuição simétrica ao redor de um valor central) que simulam as flutuações naturais das medições. A função *random.seed(42)* foi aplicada para garantir reprodutibilidade, assegurando que os mesmos dados sejam gerados em diferentes execuções do código. Em seguida, o valor de nível foi convertido para porcentagem do volume total do reservatório por meio da relação entre o nível medido e o nível máximo, multiplicada por 100. A função *clip()* da biblioteca numpy foi utilizada para limitar os valores entre 0% e 100%, evitando resultados fora dos limites físicos. Por fim, os dados foram organizados em um *data frame* da biblioteca pandas, contendo os atributos data_hora (data e hora da leitura), nivel_pct (nível em porcentagem), pressao_bar (pressão em bar) e vazao_m3_h (vazão em m³/h). Todos os valores numéricos foram arredondados para duas casas decimais com o uso da função *round()* da biblioteca numpy, garantindo maior consistência e clareza na apresentação dos resultados.

4.2. Preparação e Avaliação de Aprendizado de Máquina

O algoritmo de aprendizado de máquina empregado foi o *Random Forest Classifier*, um modelo supervisionado de classificação baseado em múltiplas árvores de decisão. Também foram utilizadas técnicas complementares de pré-processamento de dados, avaliação por métricas de desempenho, análise de importância de variáveis e interpretação de erros.

Inicialmente, o arquivo foi lido em formato de *data frame* com a função *read_csv()*, fornecida pela biblioteca pandas, possibilitando o manuseio e a análise dos dados de forma estruturada. Os dados foram tratados antes da aplicação do modelo de aprendizado. O atributo de tempo (timestamp) foi convertida para o formato de data e hora (*datetime*) e, a partir dela, foram extraídas novas variáveis temporais: hora, dia, mês e ano, o que permite ao modelo capturar possíveis padrões sazonais nas medições. Criou-se também um atributo binário chamado *status_bin*, transformando valores numéricos (0) para “normal” e (1) para “vazamento”.

Após o pré-processamento, foram selecionadas as variáveis independentes (features): nível, pressão, vazão, hora, dia e mês, que servem de entrada para o modelo. A variável dependente (*target*) escolhida foi *status_bin*, que representa o estado operacional do sistema (normal ou vazamento). Os dados foram então divididos em dois conjuntos: treinamento (80%) e teste (20%), utilizando a função *train_test_split()* do scikit-learn. Garantindo assim a avaliação em dados não vistos e obtendo um cenário realista de seu desempenho.

O algoritmo escolhido para o treinamento foi o *Random Forest Classifier*, do scikit-learn, tendo múltiplas árvores de decisão e combinando seus resultados, reduzindo o risco de overfitting. Foi configurado com 100 estimadores (árvores) e uma semente aleatória (random_state=35) para garantir reprodutibilidade. O treinamento foi realizado com o método *fit()*, utilizando os dados de entrada (*X_train*) e os rótulos correspondentes (*y_train*).

5. Desenvolvimento

O presente trabalho tem como objetivo desenvolver um modelo de aprendizado de máquina capaz de identificar automaticamente padrões anômalos no consumo de água em um reservatório, que possam indicar a presença de vazamentos. A detecção precoce desses vazamentos é essencial para reduzir desperdícios, custos operacionais e riscos ambientais.

Na Figura 6 foi criado um vetor de datas usando o *date_range()*, que gera timestamps em sequência de dados dentro do intervalo definido, por um período de dois anos, simulando 210 mil dados para que a IA possa ser treinada. Esse trecho do código gera valores simulados para nível, pressão e vazão em cada ponto da linha do tempo. Os dados são criados usando a distribuição normal, que produz números próximos da média definida, com pequenas variações realistas, mas podendo ser alterado caso for necessário para um melhor aprendizado ou teste dos resultados.

```
# 1. Parâmetros
frequencia = '5min'
hoje = datetime.now()
inicio = hoje - timedelta(days=730)  # 2 anos atrás
nivel_max = 10.0  # nível máximo do reservatório em metros

# 2. Geração do DataFrame com datas a cada 5 minutos
datas = pd.date_range(start=inicio, end=hoje, freq=frequencia)
n = len(datas)
```

Figura 6. Código de geração de parâmetros e data frame
Fonte: Autoria própria.

O código apresentado define que os dados simulados serão gerados a cada 5 minutos, capturando a data e hora atuais do sistema, que será o limite final da série temporal. Ademais, também define a data inicial como sendo 730 dias antes ou 2 anos. O nível máximo do reservatório em metros, bem como esse valor servirão como referência para converter o nível de água em percentual da capacidade total mais à frente no código. A geração do Data Frame (pd.date_range(start=inicio, end=hoje, freq=frequencia)) cria uma sequência de datas e horas que vai do início (2 anos atrás) até hoje, com intervalos de 5 minutos entre cada ponto. Isso gera todos os timestamps onde serão simuladas as leituras. Também, calcula a quantidade total de registros gerados (n = len(datas)) para 2 anos com registros a cada 5 minutos, o 'n' será em torno de 210 mil linhas.

Na sequência, a Figura 7 mostra onde o algoritmo de geração dos dados utiliza a biblioteca NumPy (np.random.normal). É usada para gerar números aleatórios com distribuição normal. Cada variável tem média (loc) e desvio padrão (scale) definidos, garantindo a variação realista em torno de valores médios esperados ou seja, cada leitura gera uma distribuição normal, garantindo variabilidade realista em torno de uma média.

```
# 3. Geração dos dados simulados
np.random.seed(42)
nivel_m = np.random.normal(loc=8.0, scale=0.3, size=n)
pressao = np.random.normal(loc=3.5, scale=0.15, size=n)
vazao = np.random.normal(loc=25.0, scale=5.0, size=n)
```

Figura 7. Código de geração de dados
Fonte: Autoria própria.

Para gerar os dados, o np.random.seed(42) define uma semente para o gerador de números aleatórios do NumPy, isso garante que toda vez que o código rodar, os mesmos valores aleatórios sejam gerados. O código apresentado na Figura 8 implementa um fluxo completo de análise e classificação de dados relacionados a medições de nível, pressão e vazão de um sistema hidráulico, com o objetivo de identificar a ocorrência de vazamentos. Inicialmente, são importadas as bibliotecas necessárias para manipulação de dados, visualização e aprendizado de máquina.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix

from google.colab import files
import io
```

Figura 8. Código de geração de dados
Fonte: Autoria própria.

Com essas bibliotecas disponíveis, o primeiro passo prático é o upload do arquivo CSV dentro do Google Colab, que contém os registros com colunas de timestamp, variáveis físicas medidas e o status do sistema se está normal ou se há vazamento.

O *upload* e leitura dos dados são feitos com a biblioteca df=pd.read_csv, demonstrado na Figura 9. Fazendo upload do arquivo .csv no Google Colab e lê os dados em um DataFrame do Pandas.

```
# 1. Upload do arquivo CSV
uploaded = files.upload()
filename = next(iter(uploaded))
df = pd.read_csv(io.BytesIO(uploaded[filename]))
```

Figura 9. Upload do arquivo
Fonte: Autoria própria.

Em seguida, realiza-se o pré-processamento dos dados (Figura 10). A coluna de tempo é convertida para o tipo datetime, permitindo a extração de hora, dia, mês e ano. Algumas colunas são renomeadas para simplificar o manuseio e o status textual é convertido em variável binária, onde 0 representa “normal” e 1 representa “vazamento”:

```
# 2. Pré-processamento
df['timestamp'] = pd.to_datetime(df['timestamp'])
df['hora'] = df['timestamp'].dt.hour
df['dia'] = df['timestamp'].dt.day
df['mes'] = df['timestamp'].dt.month
df['ano'] = df['timestamp'].dt.year  # Adiciona o ano

# Renomear colunas para facilitar (se necessário)
df.rename(columns={
    'nivel (%)': 'nivel',
    'pressao (%)': 'pressao',
    'vazao (%)': 'vazao'
}, inplace=True)

# Criar coluna binária para status
df['status_bin'] = df['status'].map({'normal': 0, 'vazamento': 1})
```

Figura 10. Pré-processamento
Fonte: Autoria própria.

Após o pré-processamento, observa-se a Figura 11, escolhendo as variáveis preditoras e a variável alvo. Nesse caso, as *features* são nível, pressão, vazão, hora, dia e mês, enquanto o ano não é usado no treino, mas será incluído mais tarde em análises adicionais.

```
# Selecionar features e target
features = ['nivel', 'pressao', 'vazao', 'hora', 'dia', 'mes']
X = df[features]
y = df['status_bin']
```

Figura 11. Features e Target
Fonte: Autoria própria.

Os dados são divididos em treino e teste, garantindo avaliação imparcial do modelo conforme a Figura 12.

```
# 3. Dividir treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figura 12. Divisão de treino e teste
Fonte: Autoria própria.

O modelo utilizado é o Random Forest. O Random Forest é um algoritmo de aprendizado de máquina, bastante utilizado para tarefas de classificação e regressão. O funcionamento básico do Random Forest parte da ideia de treinar diversas árvores de decisão independentes, esse processo envolve dois mecanismos importantes, o Bootstrap e a Seleção aleatória de atributos. E para esse modelo, foram selecionadas 100 árvores, com *random_state* de 35 (Figura 13).

```
# 4. Treinar modelo
model = RandomForestClassifier(n_estimators=100, random_state=35)
model.fit(X_train, y_train)
```

Figura 13. Treinando o modelo
Fonte: Autoria própria.

Após o treinamento com o Model Forest, o modelo realiza previsões sobre o conjunto de teste, e um relatório de classificação é exibido, contendo métricas como precisão, *recall* e f1-*score*. A precisão indica a proporção de previsões positivas que estavam corretas. O *recall* mede a proporção de casos positivos reais que foram corretamente identificados, e o fi-*score* é a média harmônica entre precisão e *recall*, equilibrando as duas métricas.

O próximo passo é avaliar se esse aprendizado é consistente quando aplicado a dados novos, que não foram vistos durante o treinamento. Para isso, utiliza-se o conjunto de teste (X_test) conforme a Figura 14. O método .predict() percorre cada linha do conjunto de teste e gera uma previsão da classe 0 para normal, 1 para vazamento. O resultado é um vetor (y_pred) que contém a classificação estimada pelo

modelo para cada registro. Esse vetor será comparado ao vetor real (y_test), que guarda as respostas corretas.

```
# 5. Avaliação
y_pred = model.predict(X_test)
print("Classification Report:\n", classification_report(y_test, y_pred))
```

Figura 14. Avaliação
Fonte: Autoria própria.

Para complementar, é gerada a matriz de confusão, na Figura 15, em forma de mapa de calor, que mostra visualmente os acertos e erros nas classificações entre as classes “normal” e “vazamento”:

```
# 6. Matriz de confusão
plt.figure(figsize=(8, 6))
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Normal', 'Vazamento'],
            yticklabels=['Normal', 'Vazamento'])


plt.xlabel("Predito")
plt.ylabel("Real")
plt.title("Matriz de Confusão")
plt.tight_layout()
plt.show()

# Espaçamento entre os gráficos
plt.figure(figsize=(8, 1))
plt.axis('off')
plt.title(" ")
```

Figura 15. Matriz de confusão
Fonte: Autoria própria.

Na Figura 16 é calculado e visualizado a importância das variáveis para o modelo. Esse gráfico de barras mostra quais atributos tiveram maior peso na detecção de vazamentos, o que ajuda a interpretar os resultados.

```
# 7. Importância das variáveis
importances = pd.Series(model.feature_importances_, index=features)
sorted_importances = importances.sort_values()

plt.figure(figsize=(8, 6))
bars = plt.barh(sorted_importances.index, sorted_importances.values, color='skyblue')
plt.title("Importância das Variáveis")
plt.xlabel("Importância")

for bar in bars:
    width = bar.get_width()
    plt.text(width + 0.005, bar.get_y() + bar.get_height()/2,
             f'{width:.3f}', va='center')

plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

Figura 16. Matriz de confusão
Fonte: Autoria própria.

Por fim, é feita uma análise detalhada dos erros do modelo. Para isso, o ano é reintroduzido no conjunto de testes, permitindo registrar em que períodos os erros ocorreram. São contabilizados falsos positivos (quando o modelo prevê vazamento mas a situação era normal) e falsos negativos (quando havia vazamento mas o modelo não detectou). Esses casos são listados para inspeção, conforme demonstrado na Figura 17.

```
# Adiciona o ano ao X_test
X_test_with_ano = X_test.copy()
X_test_with_ano['ano'] = df.loc[X_test.index, 'ano'].values

# Cria DataFrame com resultados reais e preditos
resultados = X_test_with_ano.copy()
resultados['Real'] = y_test.values
resultados['Predito'] = y_pred

# Falsos Positivos (predito vazamento, mas era normal)
falsos_positivos = resultados[(resultados['Real'] == 0) & (resultados['Predito'] == 1)]
print(f"Falsos Positivos: {len(falsos_positivos)}")
display(falsos_positivos.head())

# Falsos Negativos (predito normal, mas era vazamento)
falsos_negativos = resultados[(resultados['Real'] == 1) & (resultados['Predito'] == 0)]
print(f"Falsos Negativos: {len(falsos_negativos)}")
display(falsos_negativos.head())
```

Figura 17. Análise de falso positivo e negativo com inclusão de ano

Fonte: Autoria própria.

5. Resultados

Na classe 0, houve 40.748 amostras que representam leituras sem vazamento. Já na classe 1 há 1.358 amostras que representam leituras com vazamento. A Precision teve um acerto de 0.96, pode-se dizer que o percentual de previsões corretas dentro de cada classe, quando o modelo previu vazamento (classe 1), 96% das vezes estava correto. No Recall o percentual de casos reais são corretamente identificados e demonstra que o modelo detectou 93% dos vazamentos reais. Média harmônica de 0,95 entre Precision e Recall, equilibrando precisão e cobertura é visto no F1-Score. Há um leve impacto na classe 1, com recall de 0.93, indicando que cerca de 7% dos vazamentos reais não foram detectados. Dados vistos na Figura 17.

Observando os resultados da Figura 18, percebe-se que o modelo acertou 40.699 casos normais, classificando-os corretamente, e apenas em 49 situações gerou um alarme falso ao prever "vazamento" quando na verdade o estado era normal. Isso demonstra que o sistema possui confiabilidade em reconhecer condições normais, com baixa taxa de falsos positivos.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	40748
1	0.96	0.93	0.95	1358
accuracy			1.00	42106
macro avg	0.98	0.96	0.97	42106
weighted avg	1.00	1.00	1.00	42106

Figura 18. Classificação
Fonte: Autoria própria.

Em relação à detecção de vazamentos, o modelo identificou corretamente 1.261 ocorrências (Figura 19). No entanto, houve 97 casos de falso negativo, ou seja, situações em que havia realmente um vazamento, mas o modelo previu como normal. Pode-se ressaltar que alguns eventos críticos podem passar despercebidos.

De forma geral, o desempenho global é bom, com uma acurácia de aproximadamente 99,65%. A precisão para identificar vazamentos foi de 96,3%, o que significa que, quando o modelo aponta um vazamento, em quase todos os casos ele está correto. Já o recall, que mede a capacidade de encontrar todos os vazamentos existentes, ficou em 92,9%, mostrando que o sistema detecta a grande maioria dos casos, mas ainda pode falhar em alguns. O equilíbrio entre precisão e recall é confirmado pelo F1-score de 94,6%, considerado muito bom.

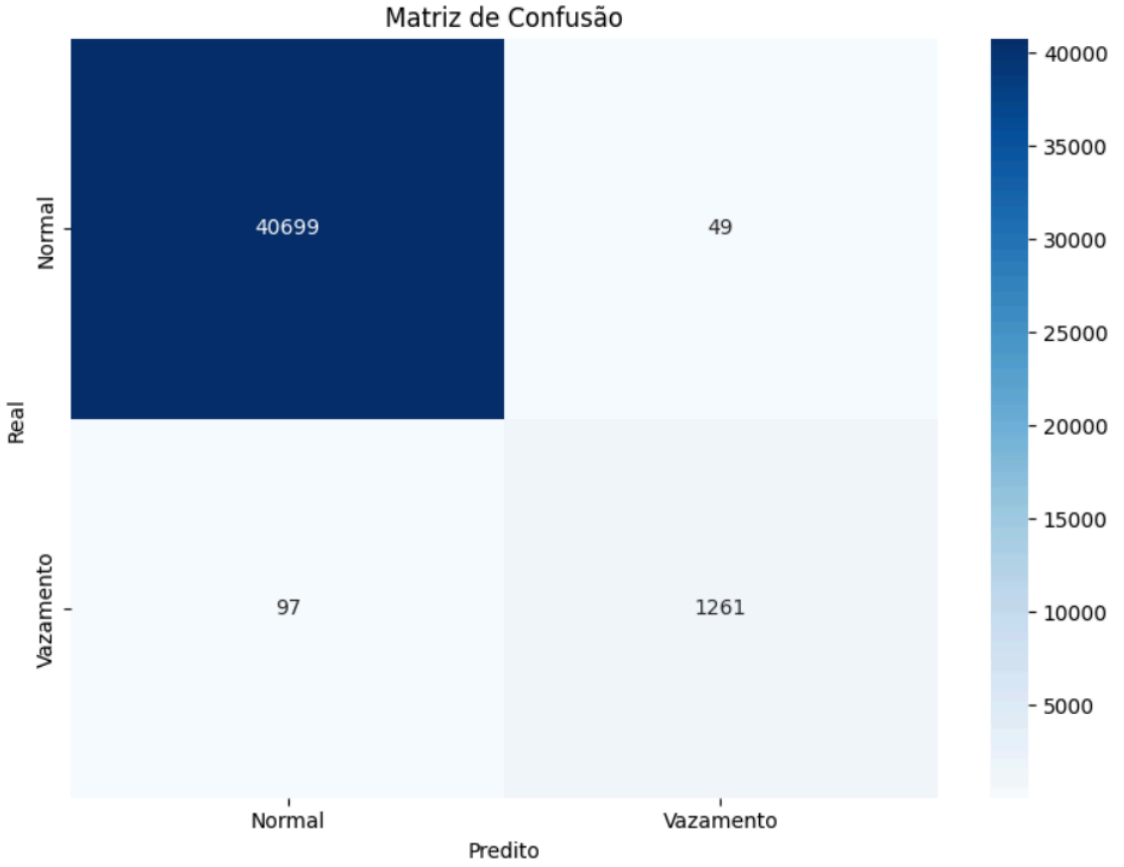

Figura 19. Matriz de confusão
Fonte: Autoria própria.

A análise da importância das variáveis, Figura 20, revelou que a vazão foi o fator mais determinante para o desempenho do modelo, respondendo por aproximadamente 48% da capacidade de classificação. Isso significa que, na maior parte dos casos, a identificação de condições normais ou de vazamento depende fortemente do comportamento da vazão, o que é coerente, já que alterações nesse parâmetro tendem a ser o primeiro indício de anomalias no sistema.

Em seguida, destacam-se as variáveis mês (15,1%), nível (14,2%) e dia (12,4%), que também tiveram relevância significativa. Esses resultados mostram que fatores temporais, como a época do ano e o dia específico, influenciam no padrão de funcionamento do sistema, possivelmente associados a sazonalidade, demanda ou condições operacionais que variam ao longo do tempo. O nível, por sua vez, aparece como variável complementar importante, reforçando a relação física entre o volume armazenado e a ocorrência de vazamentos.

Já as variáveis pressão (8,4%) e hora (1,9%) tiveram impacto menor, sendo pouco utilizadas pelo modelo para diferenciar entre normalidade e vazamento. Isso indica que, apesar de terem alguma correlação, seu poder de separação entre as classes é limitado.

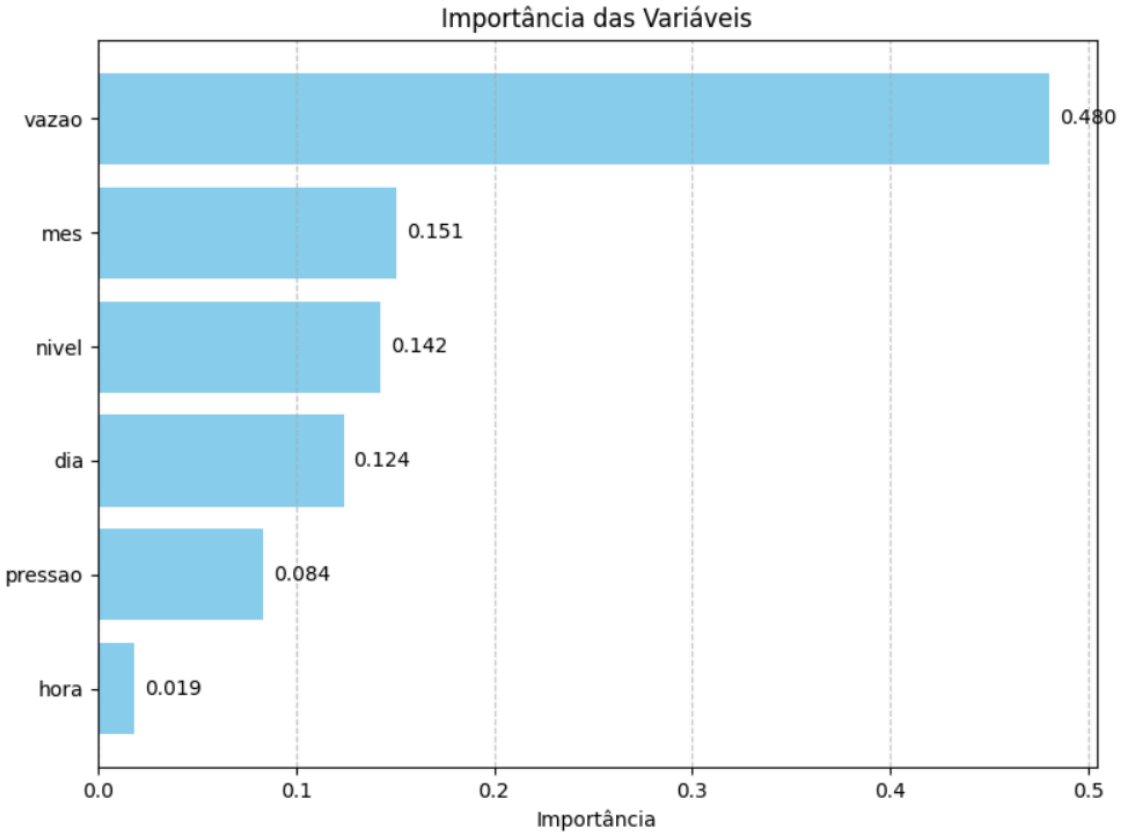

Figura 20. Importâncias das variáveis
Fonte: Autoria própria.

Na Figura 21 está visível o resultado do falso positivo e negativo. Pode-se dizer que no caso dos falsos positivos (49 ocorrências), o sistema classificou como vazamento situações que na realidade eram normais. Observa-se que, em muitos desses registros, a variável vazão apresenta valores elevados como 97.31, 95.30 ou 98.60, o que parece ter levado o modelo a interpretar incorretamente esses cenários como anômalos. Esse tipo de erro pode gerar alarmes falsos e impactar a operação, aumentando o número de verificações desnecessárias.

Já nos falsos negativos 97 ocorrências, ocorre o oposto, o modelo classificou como normal situações que eram, de fato, vazamentos. Aqui, o padrão que se destaca é a presença de vazões mais baixas ou intermediárias como 32.71, 33.50 ou 43.84 que aparentemente foram confundidas com valores típicos de funcionamento normal. Esse tipo de erro é mais crítico, pois representa a possibilidade de um vazamento e pode passar despercebido pelo sistema, gerando riscos de perdas ou falhas operacionais.

De maneira geral, os resultados evidenciam que a variável vazão desempenha um papel central nas previsões. Entretanto, também revelam uma limitação em situações de vazão muito elevada, o modelo tende a gerar falsos positivos, enquanto em níveis médios ou mais baixos, pode subestimar os problemas falsos negativos. Embora o desempenho global seja satisfatório, a análise dos erros indica a necessidade de ajustes finos no limiar de decisão ou a inclusão de novas variáveis derivadas como a relação

entre nível e pressão para aprimorar a sensibilidade na detecção de vazamentos, reduzindo especialmente os falsos negativos.

Falsos Positivos: 49

	nivel	pressao	vazao	hora	dia	mes	ano	Real	Predito
21519	13.33	11.53	97.31	17	2	10	2023	0	1
107207	31.92	28.03	78.79	5	26	7	2024	0	1
124570	11.29	10.64	95.30	12	24	9	2024	0	1
190739	69.20	62.13	39.46	6	12	5	2025	0	1
22282	14.86	13.53	98.60	8	5	10	2023	0	1

Falsos Negativos: 97

	nivel	pressao	vazao	hora	dia	mes	ano	Real	Predito
75644	74.49	69.72	32.71	15	7	4	2024	1	0
69183	77.70	69.06	33.50	5	16	3	2024	1	0
90804	61.57	57.80	43.84	7	30	5	2024	1	0
22003	17.94	16.26	92.84	9	4	10	2023	1	0
112304	23.70	20.20	73.41	22	12	8	2024	1	0

Figura 21. Falsos positivos e negativos
Fonte: Autoria própria.

6. Conclusão

O estudo apresentou resultados animadores no uso do algoritmo Random Forest para a detecção de vazamentos em sistemas de distribuição de água, demonstrando elevada acurácia e eficiência na identificação de padrões anômalos. A criação de uma base de dados foi um ponto positivo relevante, pois possibilitou a realização de testes controlados sem depender de dados reais. O modelo atingiu uma acurácia global de 99,65%, com alta precisão (96%) e recall (93%), confirmando a capacidade do sistema em reconhecer vazamentos de forma eficaz.

Entre os pontos positivos, destacam-se a robustez e interpretabilidade do algoritmo Random Forest, a metodologia clara e bem estruturada para geração e pré-processamento dos dados, e a aplicação de métricas de avaliação adequadas, que validaram o desempenho do modelo. O uso de Python e suas bibliotecas (Pandas, NumPy, Scikit-learn e Matplotlib) também reforçou a acessibilidade e replicabilidade do

estudo, permitindo que outros pesquisadores possam reproduzir e expandir os resultados.

Contudo, o trabalho também apresenta limitações. A principal delas está na utilização de uma base de dados sintética, que, embora útil para experimentação, não reflete integralmente as variações e ruídos presentes em medições reais. Além disso, o modelo apresentou falsos negativos (97 casos) e falsos positivos (49 casos), o que, em aplicações reais, pode gerar consequências práticas desde alarmes desnecessários até o não reconhecimento de vazamentos críticos. Outro ponto negativo é que o algoritmo Random Forest, apesar da boa precisão, demanda elevado tempo de processamento e pode se tornar menos eficiente em cenários com grande volume de dados em tempo real.

Como possibilidades para trabalhos futuros, recomenda-se a aplicação do modelo em dados reais coletados por sensores, permitindo avaliar sua performance em condições reais de operação. Também é altamente indicado o uso de validação cruzada estratificada, podendo ser com 10 folds, para reduzir vieses decorrentes de possíveis desequilíbrios na base e garantir que o desempenho obtido não dependa de uma única divisão dos dados. Ainda, propõe-se explorar novas features, como variabilidade temporal, pressão derivada, gradientes de vazão, entre outras variáveis, que podem enriquecer o modelo e aumentar seu poder preditivo.

Outra linha de pesquisa interessante é a integração de técnicas híbridas, combinando Random Forest com algoritmos como Redes Neurais Artificiais, Gradient Boosting ou Isolation Forest, buscando aprimorar a detecção de anomalias e reduzir especialmente os falsos negativos, que são críticos no contexto de vazamentos. Assim, o avanço dessas abordagens pode contribuir para sistemas mais precisos, confiáveis e aplicáveis em ambientes reais de monitoramento hídrico.

7. Referências

Al-Rashdan, D. *et al.* (1999) "Environmental impact assessment and ranking the environmental projects in Jordan", *European Journal of Operational Research*, 118(1), p. 30–45. Disponível em: https://doi.org/10.1016/S0377-2217(97)00079-9.

Braz, S.N. e Longo, R.M. (2021) "Qualidade ambiental das cidades: uso de bioindicadores para avalição da poluição atmosférica", *Sustentabilidade: Diálogos Interdisciplinares*, 2, p. 1–21. Disponível em: https://doi.org/10.24220/2675-7885v2e2021a5198.

Creswell, J.W. e Creswell, J.D. (2021) *Projeto de pesquisa - 2.ed.: Métodos qualitativo, quantitativo e misto*. Penso Editora. Disponível em: https://books.google.com.br/books?id=URclEAAAQBAJ&pg=PT56&hl=pt-BR&source=gbs_selected_pages&cad=1#v=onepage&q&f=false.

Favretto, C.R. *et al.* (2016) "Análise do Sistema de Abastecimento de Água do Município de Arroio do Padre/RS", em *Blucher Engineering Proceedings*. *XIV Encontro Nacional de Estudantes de Engenharia Ambiental*, Blucher Proceedings, p. 1253–1262. Disponível em: https://doi.org/10.5151/engpro-eneeamb2016-pogi-002-5063.

Flores, T.K.S. *et al.* (2021) "Medição de Vazão por Feedback Indireto em uma Rede de Bombeamento de Água Empregando Inteligencia Artificial.", *Medição de Vazão por Feedback Indireto em uma Rede de Bombeamento de Água Empregando Inteligencia Artificial.*, 21(1), p. 75. Disponível em: https://doi.org/10.3390/s21010075.

Giustolisi, O., Ridolfi, L. e Simone, A. (2019) "Tailoring Centrality Metrics for Water Distribution Networks", *Water Resources Research*, 55(3), p. 2348–2369. Disponível em: https://doi.org/10.1029/2018WR023966.

Gouveia, C.G.N. (2022) "Técnicas de aprendizado de máquina aplicadas à predição de vazamentos em ramais de redes de distribuição de água". Disponível em: http://repositorio.unb.br/handle/10482/43766 (Acesso em: 5 de outubro de 2025).

Han, J., Pei, J. e Tong, H. (2023) *Data mining: concepts and techniques*. Fourth edition. Cambridge, MA, United States: Morgan Kaufmann Publishers, an imprint of Elsevier. Disponível em: https://doi.org/10.1016/C2013-0-18660-6.

Hasan Ahmed Salman, Ali Kalakech, e Amani Steiti (2025) "Random Forest Algorithm Overview", *Random Forest Algorithm Overview*, Vol.2024, p. 69–79. Disponível em: https://doi.org/10.58496/BJML/2024/007.

Jacobi, P.R., Buckeridge, M. e Ribeiro, W.C. (2021) "Governança da Água na Região Metropolitana de São Paulo - Desafios à Luz das Mudanças Climáticas", *Governança da Água na Região Metropolitana de São Paulo - Desafios à Luz das Mudanças Climáticas*, 35, p. 209–226. Disponível em: https://doi.org/10.1590/s0103-4014.2021.35102.013.

Jesus, E. dos S. de (2023) "Modelos de aprendizagem de máquina para previsão da demanda de água da região metropolitana de Salvador, Bahia." Disponível em: https://doi.org/10.1007/s00521-023-08842-0.

Kanakoudis e Muhammetoglu (2014) "(PDF) Urban Water Pipe Networks Management Towards Non-Revenue Water Reduction: Two Case Studies from Greece and Turkey", *Urban Water Pipe Networks Management Towards Non-Revenue Water Reduction*, 42. Disponível em: https://doi.org/10.1002/clen.201300138.

Kelleher, J.D., MacNamee, B. e D'Arcy, A. (2015) *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. Cambridge, Massachusetts London, England: The MIT Press.

Nusda, F.F. (2023) "Proposta de Modelo para Reconhecimento de Padrões de Comportamento de Vazamentos de Água Visando a Redução de Perdas na Distribuição", *Proposta de Modelo para Reconhecimento de Padrões de Comportamento de Vazamentos de Água Visando a Redução de Perdas na Distribuição*, p. 138. Disponível em: https://repositorio.utfpr.edu.br/jspui/bitstream/1/32247/1/modelovazamentoreducaoperdas.pdf.

Pedregosa, F. *et al.* (2011) "Scikit-learn: Machine Learning in Python", *J. Mach. Learn. Res.*, 12, p. 2825–2830. Disponível em: https://doi.org/10.5555/1953048.2078195.

Pinnto, M.R. *et al.* (2017) “Dimensionamento econômico otimizado de redes de distribuição de água considerando custos de manutenção”, *Engenharia Sanitaria e Ambiental*, 22, p. 145–153. Disponível em: https://doi.org/10.1590/S1413-41522016140349.

Porto, R.M. (2006) “Hidráulica Básica”. EESC/USP. Disponível em: https://zonadaeletrica.com.br/hidraulica-basica-4a-edicao-rodrigo-de-melo-porto/ (Acesso em: 5 de outubro de 2025).

Russell, Stuart e Norvig, Peter (2011) *Inteligência artificial*. 3rd ed. Rio de Janeiro: Elsevier.

SANSA (2022) *Diagnostico Temático Serviços De Água E Esgoto*. Disponível em: https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/saneamento/snis/produtos-do-snis/diagnosticos/DIAGNOSTICO_TEMATICO_VISAO_GERAL_AE_SNIS_2023.pdf (Acesso em: 5 de outubro de 2025).

Silva, D. do R. e (2025) “Utilização da Inteligência Artificial para Redução de Perdas de Água em Sistemas de Abastecimento”, *Utilização da Inteligência Artificial para Redução de Perdas de Água em Sistemas de Abastecimento* [Preprint]. Disponível em: https://ric.cps.sp.gov.br/bitstream/123456789/35549/1/analiseedesenvolvimentodesistemas_2025_1_danilodorosarioesilva_utiliza%c3%a7%c3%a3odaintelig%c3%aanciaartificialpararedu%c3%a7%c3%a3o.pdf (Acesso em: 5 de outubro de 2025).

Tsutiya, M.T. (2006) *Abastecimento de água*. 3. ed. São Paulo (SP): Escola Politécnica da Universidade de São Paulo. Disponível em: https://pt.scribd.com/doc/127473795/Abastecimento-de-Agua-Tsutiya.